Chapter 100
A Semi-Supervised Network Traffic Classification Method Based on Incremental Learning

Pinghong Li, Yong Wang and Xiaoling Tao

Abstract In order to solve low accuracy, time consumption and limited application range in traditional network traffic classification, a semi-supervised network traffic classification method based on incremental learning is proposed. During training Support Vector Machine (SVM), it takes full advantage of a large number of unlabeled samples and a small amount of labeled samples to modify the classifiers. By utilizing incremental learning technology to void unnecessary repetition training, improve the situation of original classifiers' low accuracy and time-consuming when new samples are added. Combined with the Synergies of multiple classifiers, this paper proposes an improved Tri-training method to train multiple classifiers, overcoming the strict limitation of traditional Co-verification for classification methods and sample types. Experiments' results show that the proposed algorithm has excellent accuracy and speed in traffic classification.

Keywords Traffic classification · Support vector machine · Semi-supervised · Incremental learning · Tri-training

100.1 Introduction

Network traffic is an important carrier of recording, reflecting the network status and user activities, it plays an increasingly important role in effective network management. Network traffic classification [1] classifies the two-way TCP or UDP

P. Li (✉) · Y. Wang · X. Tao
Guilin University of Electronic Technology, NO.1 Jin-ji Road, Qixing District, Guilin, Guangxi, China
e-mail: lipinghong0601@163.com

W. Lu et al. (eds.), *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, Lecture Notes in Electrical Engineering 211,
DOI: 10.1007/978-3-642-34522-7_100,

stream generated by network communication according to the types of network applications (such as WWW, FTP, MAIL, P2P) in the Internet based on TCP/IP protocol.

Recently, applying machine learning method to classify and identify network applications is a research hotspot. There are two traditional strategies in machine learning [2], that's supervised learning and unsupervised learning. Supervised learning methods, such as Bayesian methods, Decision tree methods, are high detection rates, but require that the sample data is correctly marked in advance and they are unable to find the unknown category samples. Unsupervised learning methods, such as Clustering method, group samples according to the data similarity. They don't need labeled data, but only model unlabeled data, detection accuracy is low.

Semi-supervised learning can take full advantage of a large number of unlabeled samples and a small amount of labeled samples. It makes up for the shortage of supervised learning and unsupervised learning. In this paper, a novel Least Area-SVM (LA-SVM) traffic classification algorithm is proposed, and we use improved Tri-training method to train classifiers collaboratively based on semi-supervised learning, the method makes the most of incremental learning in the classification efficiency and collaborative training techniques in accuracy, which improves network traffic classification performance.

100.2 SVM and Incremental Learning

100.2.1 SVM

SVM is an efficient and general machine learning algorithm based on Statistical Learning Theory (SLT). It's goal is to separate two classes by Constructing an objective function. Compared with conventional machine learning methods, SVM has many advantages [3]. (1) Global optimal solution. (2) Good Generalization performance. (3) Kernel skills application. (4) Good robustness [3, 4].

For the classification problems, if sample set is $\{X_i, Y_i\}$, i = 1,…,l. $X_i \in R^n$, $Y_i \in \{-1, +1\}$, Maximizing the distance of the hyper plane with the nearest samples to ensure the classification accuracy. If the classification problem is nonlinear, the input space is mapped into high dimensional feature space by using kernel functions. When and only if each support vector a satisfies the KKT conditions, $a = [a_1, a_2, \ldots, a_l]$ is the optimal solution. The KKT conditions as formula (100.1)

$$\begin{cases} a_i = 0 \Rightarrow f(x_i) \geq 1 \; or \; f(x_i) \leq -1 \\ 0 < a_i < C \Rightarrow f(x_i) = 1 \; or \; f(x_i) = -1 \\ a_i = C \Rightarrow -1 \leq f(x_i) \leq 1 \end{cases} \tag{100.1}$$

where a is Lagrange multiplier, when $a > 0$, the corresponding samples are called Support vector. C is the regularization parameter.

100.2.2 Incremental Learning

With the development of modern technologies, the ever-growing network traffic information is increasingly large, it is very difficult to obtain a complete training data set at an early stage. This requires the classifiers can continuously improve the learning accuracy with the accumulation of data samples, so the incremental learning is very important.

For standard SVM Incremental learning algorithm, it takes support vector set obtained from last training as historical learning results instead of training samples during the training of SVM. Batch SVM [5] incremental learning method divides new samples into several disjoint subsets, and gets the final results by constructing new support vector set and classification hyper plane Sequence, but these serial incremental learning strategies can not reduce the time complexity of the classification process [6]. The two typical incremental learning algorithms do not fully consider the initial samples and new samples which may be converted to support vector data, leading to some useful historical data to be eliminated early and affecting the classification accuracy.

100.3 Semi-Supervised Learning and Co-Training

Co-training is a kind of semi-supervised learning paradigm that was proposed by Blum and Mitchell first [7]. It assumes that attributes can be split into two sufficient and redundant views. Tow independent classifiers are trained with the labeled data. Then each classifier labels unlabeled data with samples of high confidence that are selected from unlabeled data, and puts them into the labeled training set in the other classifier.

As most data can't meet fully redundancy conditions of views, Goldman and Zhou proposed an improved Co-training algorithm [8]. It no longer requires the problem itself has fully redundant views, but 10 times cross-validation to determine unlabeled samples' confidence. Its disadvantage is time-consuming. Zhou proposed Tri-training algorithm [9] for solving the problem. It uses three classifiers and doesn't require the data be described with sufficient and redundant views, but its auxiliary classifiers may produce noise samples. In addition, Deng [10] proposed adaptive data editing algorithm based on Tri-training and proved it's fast and easy to extend for common data.

100.4 SVM Network Traffic Classification with Incremental Learning and Improved Tri-Training

100.4.1 Changes of Support Vectors After Adding New Samples

Zhou [11] proved new samples which meet KKT conditions will not change support vector sets, while new samples against KKT conditions do. Wang [12] proved if there are new samples against the KKT conditions, non-support vectors in original samples may be converted to support vectors. It can be concluded that: the classifier's performance depends mainly on samples in new samples against the KKT conditions, support vector set in original samples and non-support vectors which may be converted to support vectors in original samples.

In this paper, Fig. 100.1 gives the illustration on Changes of support vectors after adding new samples. A1, A2, A3, A4 are support vectors in original samples, B1, B2, B3, B4 are newly added samples, when they are added, the support vectors become C1, C2, C3, C4, B1, B3, A1. Among them, C1, C2, C3, C4 are non-support vectors in original samples, B1, B3 are new samples, A1 is support vector in original samples. This show that support vectors changes when new samples are added. Firstly, non-support vectors C1, C2, C3, C4 in original samples and new samples B1 and B3 become support vectors. Then, although A1, A2, A3, A4 are support vectors in original samples, only A1 becomes new support vector after new samples are added. Therefore, how to ensure support vectors is of importance.

100.4.2 LA-SVM Method

According to the analysis above, for SVM incremental learning in the application of incremental learning, the traditional algorithms mostly not fully consider the initial samples and new samples which can be transformed into support vectors, which results in some useful data are prematurely eliminated and affecting classification accuracy. The starting point of LA-SVM method is finding out non-support vectors which can be transformed into new support vectors in the original sample sets.

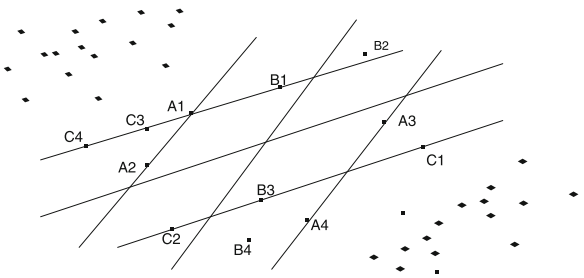

Fig. 100.1 Changes of support vectors after adding new samples

For new samples, LA-SVM method only keeps samples against KKT conditions. For initial samples, LA-SVM method only keeps support vectors, and non-support vectors within $[-1-u,\ 1+u]$ (u is Threshold) area from support vectors.

Based on the ideas above, LA-SVM method can be described as: Assume the initial sample set is X_0, the new sample set is X_i, i = 1,…, n, and $X_0 \cap X_i = \Phi$. The purpose is to find new classifier based on $X_0 \cap X_i$ and the corresponding support vector set SV. The specific steps as follows:

Step 1: train initial sample set X_0 to get SVM classifier T_0, Support vector set X_0^{sv} and non- Support vector set X_0^{nsv} of the initial sample.
Step 2: verify samples in X_i to check whether there are samples against the KKT condition, if so, divide X_i into X_i^{sv} (samples that satisfy the KKT condition) and X_i^{nsv} (samples that don't satisfy the KKT condition). If not, jump to step 4.
Step 3: search sample points within$[-1-u,\ 1+u]$ away from X_0^{sv}, collect them as set X_a, delete repeated points in X_a, and define the rest points in X_a as X_A.
Step 4: set $X = X_0^{sv} \cup X_A \cup X_i^{nsv}$, get SVM classifier T and support vectors SV by training X.
Step 5: T and SV are what we need.

100.4.3 Semi-Supervised SVM Based on Improved Tri-Training

The improved Tri-training method needs three learners as classifiers. It has no special demand for these classifiers. Let X_0 denote the initial labeled example set, X_u denote the unlabeled. The specific steps as follows:

Step 1: train labeled sample sets by bootstrap sampling to obtain three labeled training sets. By training the three labeled training sets in LA-SVM algorithm to achieve there initial classifiers A, B and C.
Step 2: after the initial training, one of the three classifiers will act as the training target classifier (assume it is A) and the others are auxiliary classifiers (assume they are B and C).
Step 3: B and C are used to classify samples in set X_u, if they have reach a consensus on the label of an unlabeled sample, the sample and the corresponding labels will be gathered together as $X_{a'}$.
Step 4: let $U = X_0 U X_{a'}$, retrain classifier A by set U to get $X_{A'}$, and put $X_{a'}$ back to unlabeled set X_u.
Step 5: compare $X_{A'}$ with previous classifier, if there are changes, jump to step 3; if there are no changes, jump to Step 6.
Step 6: Training end. Classifier $X_{A'}$ is what we need.

For improved Semi-supervised Tri-training method, it should be noted that $X_{a'}$ is not as the labeled data and will be put back to unlabeled set X_u in next round. If

$X_{a'}$ is correct prediction, the training target classifier will have additional correct samples. If $X_{a'}$ is wrong prediction, it will get extra nosy samples [10]. The noise will decrease the classifiers' performance, that's why this paper lets $X_{a'}$ as the unlabeled sample in the next round. By this way, it can reduce the classification error rate. The method has no constraint for attribute sets and semi-supervised learning algorithms used in three classifiers. Without cross-validation, therefore it is applicable to a much wider range and more high efficient.

100.5 Experiments

In experiments, we select data set used in paper [13] by Professor Moore, computer department of Cambridge University, this paper called it Moore-set. The Moore-set is the most authoritative network traffic classification test data set, which provides 10 traffic classification data subsets, each subset contains tens of thousands of data. Each stream contains 249 properties, which are composed of the statistical properties such as the port number and network properties such as the average time interval. The last property is classification target attribute, indicating types of the traffic samples, such as the WWW, FTP, P2P and so on. All our experiments run on Windows XP with MATLAB V7.1 and Myeclipse V8.5.0 installed.

100.5.1 Experiments Results

The experiment extracted 10 % of each subset in Moore_set, a total of 37,740 network traffic data. In order to verify the ability of proposed LA-SVM to process labeled and unlabeled samples in the same time under Tri-training. Compare training time and accuracy of standard SVM, Batch SVM and LA-SVM incremental learning under the unlabeled rate from 80 to 25 %. The experimental results are shown in Figs. 100.2 and 100.3. In order to verify the incremental learning ability of proposed LA-SVM under Tri- training, Select 40 % of samples as the initial sample set, divide the rest of the sample into 11 parts and add one part once. Compare training time and accuracy of standard SVM, Batch SVM and LA-SVM incremental learning. The experimental results are shown in Figs. 100.4 and 100.5.

100.5.2 Experiments Analysis

Figures 100.2 and 100.3 indicate the variation tendency of classification accuracy and training time under different unlabeled rates. Figure 100.2 shows that the

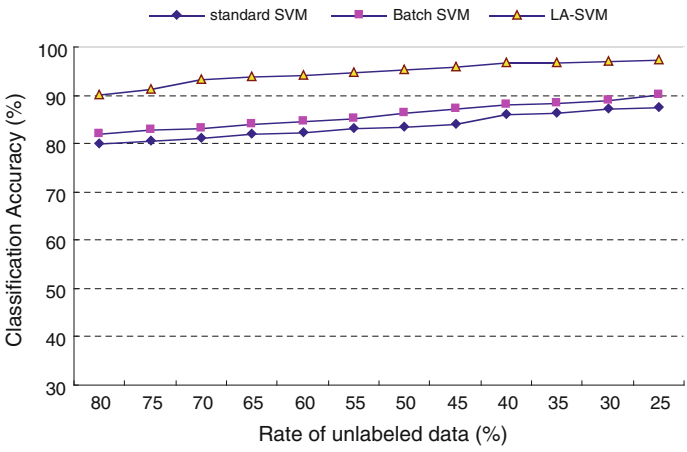

Fig. 100.2 Accuracy under different unlabeled rates

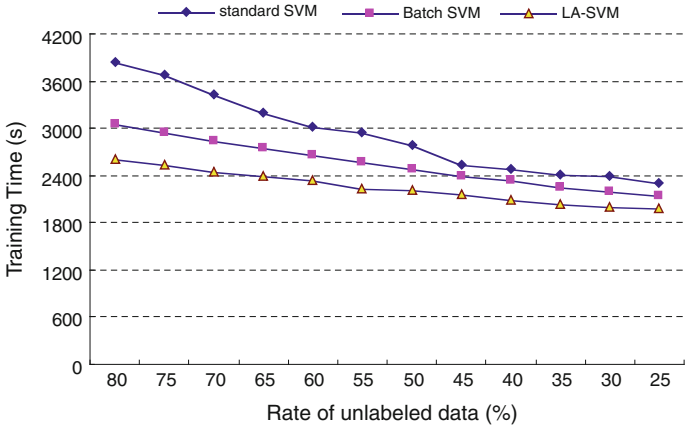

Fig. 100.3 Time under different unlabeled rates

classification accuracy of LA-SVM is better than standard SVM and Batch SVM incremental learning at different stages. The accuracy difference is the maximum under 80–70 % unlabeled rate. But classification accuracy rate of the three methods increase not obviously under 20–30 % unlabeled rate. Figure 100.3 reveals that the training time of LA-SVM costs less than the other two algorithms under different unlabeled rates. That's because LA-SVM incremental learning can

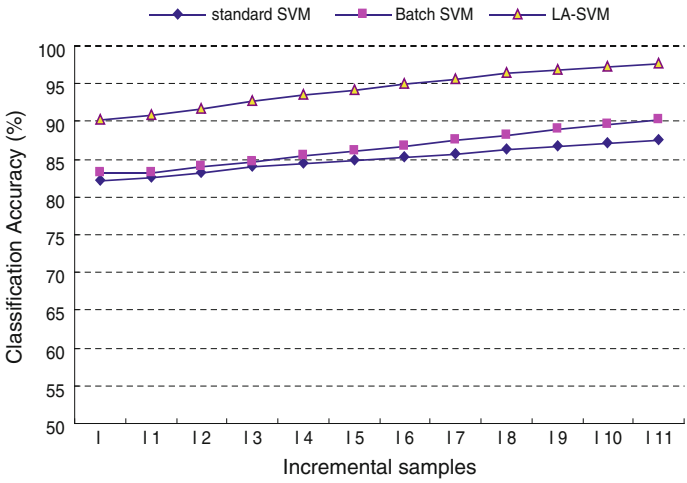

Fig. 100.4 Accuracy with incremental samples

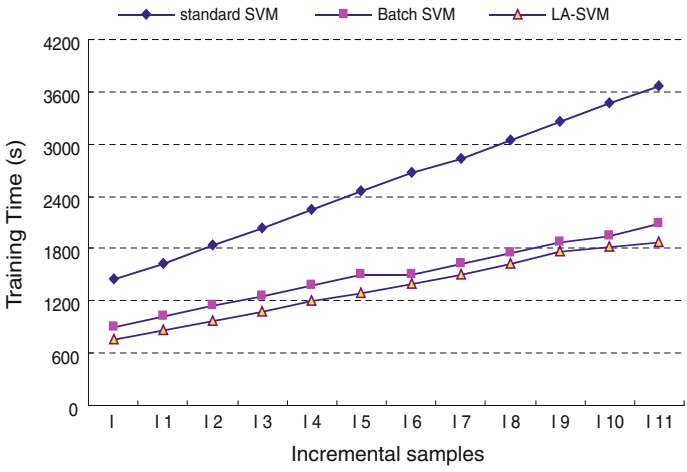

Fig. 100.5 Time with incremental samples

take full advantage of a large number of unlabeled samples and a small amount of labeled samples.

Figures 100.4 and 100.5 indicate the variation tendency of classification accuracy and training time with incremental samples, and we can see that the performance of LA-SVM in classification accuracy and time-consuming are much better than the other two algorithms with new samples are continuously added. The reason in that LA-SVM incremental learning adopts reasonable historical data elimination mechanism, which Consider the influences of Non-support vector in

the initial sample set for incremental learning. It achieved high accuracy and efficiency. The experiments show that the LA-SVM algorithm is feasible and effective.

100.6 Conclusion

In this paper, a novel LA-SVM incremental learning algorithm is proposed. Compared with other methods, it fully consider samples in new samples against the KKT conditions, support vector set in original samples and non-support vectors in original samples which may be converted to support vectors. It avoids useful data to be eliminated early. Incremental learning techniques make full use of the results of historical study, significantly reducing training time. Finally, we use improved Tri-training method to train the classifiers. Experiments show that the method proposed has preponderance in network traffic classification accuracy and speed.

The methods of Support Vector Machine and Incremental Learning have bright prospects in the network traffic classification. Recently, researchers proposed some solutions and improved the algorithms from different aspects, but there are still some problems need to study. Firstly, how to collect valuable samples as little as possible. Secondly, how to realize fuzzy SVM incremental learning. Lastly, how to achieve online incremental learning. Due to traffic classification data is large scale, research on incremental learning algorithm with multiple support vector machine classifiers may be a research direction for traffic classification.

References

1. Zhang B, Yang J-H, Wu J-P (2011) Survey and analysis on the internet traffic model. J Softw 22(1):115–131 (in Chinese)
2. Tan PN, Steinbach M, Kumar V (2006) Introduction to data mining. Addison Wesley, Boston
3. Zhu XJ Semi-supervised learning literature survey, Technical Report 1530. Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2007–12
4. Gu C, Zhang S (2011) Network traffic classification based on improved support vector machine. Chin J Sci Instrum 32(7):1507–1513
5. Ratnasamy S, Francis P, Handley M et al (2001) A scalable content-addressable network. ACM SIGCOMM, San Diego
6. Rowstron A, Druschel P (2001) Pastry: scalable, decentralized object location, and routing for large-scale peer-to-peer system. ACM IFIP international conference on distributed systems platforms (Middleware 2001), Heidelberg, 2001
7. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training/ proceedings of the 11th annual conference on computational learning theory, Madison, 1998, pp 92–100
8. Goldman S, Zhon Y (2000) Enhancins supervised learning with unlabeled data/proceedings of the17th IGML. Morffan Kaufmann, San Francisco, GA, pp 327–334

9. Zhou ZH, Li M (2005) Tri-training: exploiting unlabeled data using three classifiers. IEEE Trans Knowl Data Eng 17(11):1529–1541
10. Deng C, Guo M-Z (2007) Tri-training with adaptive data editing. Chin J Comput 30(8):1214–1226 (in Chinese)
11. Zhou W-D, Zhang L, Jiao L-C (2001) An improved principle for measuring generalization performance. Chin J Comput 29(5):590–594 (in Chinese)
12. Wang X-D, Zheng C-Y, Wu C-M, Zhang H-D (2006) New algorithm for SVM-Based incremental learning. J Comput Appl 26(10):2440–2443 (in Chinese)
13. Moore AW, Zuev D (2005) Internet traffic classification using Bayesian analysis techniques. In: Proceedings of the 2005 ACM SIGMETRICS international conference on measurement and modeling of computer systems, pp 50–60